11-2016

On the comparison of the strength of morphological integration across morphometric datasets

Dean C. Adams
Iowa State University, dcadams@iastate.edu

Michael L. Collyer
Western Kentucky University

Brief Communication

ON THE COMPARISON OF THE STRENGTH OF MORPHOLOGICAL INTEGRATION ACROSS MORPHOMETRIC DATASETS

Dean C. Adams[1,3] and Michael L. Collyer[2]

[1]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames IA, USA*

Department of Statistics, Iowa State University, Ames IA, USA

[2]*Department of Biology, Western Kentucky University, Bowling Green, KY, USA*

[3]*Corresponding author email: dcadams@iastate.edu*

Short title: Comparing integration strength across datasets

This is the peer reviewed version of the following article: Adams, D. C. and Collyer, M. L. (2016), On the comparison of the strength of morphological integration across morphometric datasets. *Evolution*, 70: 2623–2631, which has been published in final form at doi:10.1111/evo.13045. This article may be used for non-commercial purposes in accordance With Wiley Terms and Conditions for self-archiving.

Abstract

Evolutionary morphologists frequently wish to understand the extent to which organisms are integrated, and whether the strength of morphological integration among subsets of phenotypic variables differ among taxa or other groups. However, comparisons of the strength of integration across datasets are difficult, in part because the summary measures that characterize these patterns (RV and r_{PLS}) are dependent both on sample size and on the number of variables. As a solution to this issue we propose a standardized test statistic (a *z*-score) for measuring the degree of morphological integration between sets of variables. The approach is based on a partial least squares analysis of trait covariation, and its permutation-based sampling distribution. Under the null hypothesis of a random association of variables, the method displays a constant expected value and confidence intervals for datasets of differing sample sizes and variable number, thereby providing a consistent measure of integration suitable for comparisons across datasets. A two-sample test is also proposed to statistically determine whether levels of integration differ between datasets, and an empirical example examining cranial shape integration in Mediterranean wall lizards illustrates its use. Some extensions of the procedure are also discussed.

Introduction

Over the past several decades, evaluating the degree to which morphological traits covary (*morphological integration*: sensu Olson and Miller 1958), has become a prominent subject in evolutionary biology (Cheverud 1982; Cheverud 1996; Bookstein et al. 2003; Pigliucci 2003; Klingenberg 2008; Goswami and Polly 2010). Myriad studies have characterized patterns of morphological integration in a variety of organisms, in an effort to decipher the genetic, developmental, and functional mechanisms that generate such patterns (e.g., Hallgrímsson et al. 2002; Mitteroecker et al. 2004; Monteiro et al. 2005; Young and Badyaev 2006; Gómez-Robles and Polly 2012). These empirical studies have been facilitated in part by the development of quantitative approaches for characterizing patterns of morphological integration in high-dimensional data (e.g., Magwene 2001; Bookstein et al. 2003; Mitteroecker and Bookstein 2007; Márquez 2008; Klingenberg 2009; Adams and Felice 2014; Bookstein 2015). In particular, methods that evaluate covariance patterns across *a priori* subsets of variables have received considerable attention.

In the field of geometric morphometrics, a number of approaches are utilized for characterizing the integration among subsets of variables. One approach is based on Escoffier's (1973) RV coefficient (Klingenberg 2009), which is a ratio describing the degree of covariation between sets of variables relative to the variation and covariation within sets of variables. The RV coefficient ranges between zero and one, and larger values describe greater covariation between sets of variables relative to within them, which may provide evidence that there is higher integration among subsets than expected by chance. An alternative measure is based on partial least squares (PLS), where a singular value decomposition of the covariance matrix between two sets of variables ($\mathbf{S_{12}}$) is used to describe the maximal covariation between them (Bookstein et al. 2003; Mitteroecker and Bookstein 2007). The dominant singular value of $\mathbf{S_{12}}$ explains the maximal covariation between the two sets of variables, whose pattern of covariation is described by the first set of linear combinations (singular vectors) in each of the two datasets (Bookstein et al. 2003). Scores projected on these axes are routinely used to estimate the maximal correlation among sets of variables (r_{PLS}: Rohlf and Corti 2000), with higher correlations indicating a greater level of

covariation. For both the RV and PLS approaches, statistical evaluation of the observed pattern is accomplished using permutation, where the rows (individuals) are shuffled in one subset of variables while leaving the rows in the other subset constant, thereby disassociating the covariation between subsets and generating a distribution of possible outcomes under the null hypothesis of no association between variable subsets. The observed statistic is then compared to a distribution of random statistics obtained from this procedure to evaluate its significance (see Rohlf and Corti 2000; Bookstein et al. 2003; Klingenberg 2009).

Recently, there has been increased interest in understanding the extent to which patterns of morphological integration are consistent across levels of biological organization, and whether levels of integration change over evolutionary time (Armbruster et al. 2014; Goswami et al. 2014; Klingenberg 2014). To this end researchers have characterized levels of integration across traits and species using one or more of the methods mentioned above for subsequent qualitative or quantitative comparison (for recent examples see: Drake and Klingenberg 2010; Goswami et al. 2014; Lazic et al. 2015; Martin-Serra et al. 2015; Neaux et al. 2015). However, for such comparisons to be meaningful requires that the evaluated test measures are unaffected by other attributes of the data. Unfortunately this is not the case. For instance, the RV coefficient has been shown to be sensitive to both the sample size (n) and the number of variables examined (p), rendering comparisons of RV measures across datasets uninformative (Adams 2016; also: Smilde et al. 2009; Fruciano et al. 2013; for an extended critique of the RV coefficient see: Bookstein 2016). Additionally, as shown in part by Mitteroecker and Bookstein (2007), and comprehensively below, the PLS correlation coefficient (r_{PLS}) suffers from the same inherent issues. The objective of the current manuscript is to provide a standardized test statistic (a z-score) for measuring the degree of morphological integration between sets of variables. Our procedure is developed for and is used on a PLS correlation of among-partition trait covariation. However, it is sufficiently flexible that it may be applied to any meaningful measure that captures the degree of integration in a dataset, and is thus a useful approach for comparing the degree of integration as new analytical approaches are developed (see Discussion).

Sample Size and Variable Dependency of r_{PLS}

To understand the properties of r_{PLS} we conducted simulations similar to those of Adams (2016). Specifically, simulated datasets were obtained by generating random variables drawn from a normal distribution ~$N(0,1)$, and variables were randomly assigned to one of two subsets with the constraint that the number of variables was the same in each subset. Thus, each simulated dataset represented what was expected under the null hypothesis of a random association of variables. Using this procedure, we generated 100 datasets for differing levels of sample size (n), where the total number of variables was the same (p = 30). Next we performed the reciprocal simulation where all datasets contained the same number of specimens (n = 100), but where the total number of variables differed. From each simulated dataset r_{PLS} was estimated, and at each level of n and p, the mean and 95% confidence intervals across the 100 datasets were calculated. All simulations were performed in R 3.2.0 (R Core Team 2015).

As is clear from Figure 1, values of r_{PLS} vary between zero and one, with larger values attained under smaller sample sizes, as well as with a larger number of variables (Fig. 1A, B: see also Mitteroecker and Bookstein 2007). Thus, like the RV coefficient, estimates of morphological integration using partial least squares are also sensitive to n and p, rendering comparisons of these values across datasets challenging (see also Mitteroecker and Bookstein 2007). Thus, for this purpose an alternative estimate of the degree of integration across sets of variables is required.

The *Z*-Score for Comparing the Strength of Integration

Although studies of integration rarely state a null hypothesis in terms of the parameters tested, the permutation-based procedure described above evaluates the observed measure against a distribution of values obtained under a null hypothesis of no association between subsets of variables. Thus, some generalization of the Pearson product-moment null hypothesis, $\rho = 0$, could be implied. However, whereas Pearson's r, as an estimate of ρ, has an expected value of 0 under the null hypothesis for univariate tests, r_{PLS} has a lower limit of 0 and an expected value that varies with n and p (Fig. 1 A, B).

For single-sample hypotheses, stating the null hypothesis as "no association" between matrices is sufficient; the expected value is simply the mean of the sampling distribution of r_{PLS} from the permutation procedure (as described above) and the percentile of the observed r_{PLS} value is the estimate of the *P*-value. To either qualitatively compare or actually test the dissimilarity of two measures of integration from two samples requires calculation of effect sizes in relation to expected values under the null hypothesis of no integration (e.g., Collyer et al. 2015), especially if the two samples have different expected values. This can be accomplished by calculating the standard deviates (effect sizes) of r_{PLS} for the different samples,

$$\hat{Z} = \frac{r - \hat{\mu}_r}{\hat{\sigma}_r}, \tag{1}$$

where $\hat{\mu}_r$ is the estimated expected value of r_{PLS} under the null hypothesis, found as the mean of the sampling distribution, and $\hat{\sigma}_r$ is the standard deviation of the sampling distribution (i.e., standard error of the mean). Calculating effect sizes this way assumes that the sampling distribution is normally distributed, a property we demonstrate via simulation (below).

At first glance, the numerator of the standard deviate calculation might also seem sufficient for calculating a statistic that allows integration to be compared between sets of variables. Indeed, the numerator of the standard deviate calculation detrends r_{PLS} values that might have different expectations under the null hypothesis (Fig. 1 C, D). However, there are two important concerns with using detrended r_{PLS} as a comparable statistic. First, although the statistic is no longer *n*- or *p*-dependent, the standard error (and thus, confidence interval) of the statistic varies across *n* and *p* for the same number of random permutations, decreasing with either increased *n* or *p* (Fig. 1 C, D). Second, the value itself has little meaning as a correlation coefficient. For example, negative values do not indicate negative covariation, but rather less covariation than expected under the null hypothesis. Furthermore, because low *n* or high *p* can produce large expected r_{PLS} values (Fig. 1 A, B), a veritable strong correlation between two sets of variables with large sample size will have a small detrended value, as a maximum r_{PLS} value of 1.0 precludes large detrended values for large samples. Therefore, standardization – dividing by the standard deviation of the sampling (null) distribution – is needed to generate a test statistic (*z*) that has a constant

expected value and same variance, standard deviation, or confidence interval across the entire spectrum of sample sizes and variable number (Fig 1. E, F).

We wish to reiterate that other than concern for *n*- or *p*-dependency for comparisons of integration across multiple datasets, r_{PLS} is sufficient as a single-sample test statistic. The effect size calculation in Equation 1 is merely descriptive, but allows qualitative comparison between two measures of integration that have different expected values. (Inferring the probability of the effect size from a standard normal distribution is not necessary, as it is already accomplished from the resampling experiment of the PLS analysis, as shown below.) For a statistical treatment of the comparisons of two or more datasets, however, Equation 1 can be modified to calculate the effect size of the difference between two integration effect sizes as

$$\hat{z}_{12} = \frac{\left|\left(r_1 - \hat{\mu}_1\right) - \left(r_2 - \hat{\mu}_2\right)\right|}{\sqrt{\hat{\sigma}^2_{r_1} + \hat{\sigma}^2_{r_2}}}. \quad (2)$$

The probability of $\hat{z}_{12}$ under the null hypothesis of equal integration (*P*-value) can be estimated from a standard normal distribution, provided the within-sample random r_{PLS} values are approximately normally distributed. The numerator of Equation 1 can be rewritten as $\left(r_1 - r_2\right) - \left(\hat{\mu}_1 - \hat{\mu}_2\right)$, which characterizes the effect size as a difference in levels of integration compared to their expected difference. We assert that a null hypothesis test for comparing levels of integration is naturally a two-tailed test (hence the absolute value in the numerator of Equation 2), as the direction of the difference between detrended r_{PLS} scores has no appreciable meaning, especially in the absence of standardization. For example, a smaller value of $\left(r_2 - \hat{\mu}_2\right)$ than $\left(r_1 - \hat{\mu}_1\right)$ might produce a positive difference, $\left(r_1 - \hat{\mu}_1\right) - \left(r_2 - \hat{\mu}_2\right)$, even if $\hat{z}_2$ is larger than $\hat{z}_1$, because of a smaller standard error in the second sample (perhaps because of a larger expected value from the sampling distribution). Therefore, only the magnitude of the effect size of the integration difference is valuable as a test statistic, rendering the hypothesis inherently two-tailed. Likewise, a

157 confidence interval for the effect size can be estimated as

$$(1-\alpha)100\%CI = \left|(r_1-\hat{\mu}_1)-(r_2-\hat{\mu}_2)\right| \pm z_{\alpha/2}\sqrt{\hat{\sigma}^2_{r_1}+\hat{\sigma}^2_{r_2}}, \quad (3)$$

158

159 where $z_{\alpha/2}$ is the quantile from a standard normal distribution corresponding to the two-tailed probability

160 for the level of significance, α, and $\sqrt{\hat{\sigma}^2_{r_1}+\hat{\sigma}^2_{r_2}}$ is the pooled standard error. The null hypothesis is

161 rejected with $(1-\alpha)100\%$ confidence, if the confidence interval does not contain 0, the expected

162 difference in effect sizes under the null hypothesis. Thus, concerning hypothesis tests, confidence

163 intervals and the two-sample *z*-score are test statistics for the difference in r_{PLS} between different data sets,

164 relative to their expected difference based on *n*- or *p*-dependency. The effect sizes (equation 1) allow one

165 to infer which sample has greater morphological integration, when a significant difference is observed.

166

167 **Statistical Properties**

168 To determine if this test of difference in the strength of integration among samples has

169 appropriate statistical properties (type I and type II error rates, statistical power), we performed a

170 simulation experiment. This experiment consisted of 100 runs of generating two "populations" of data,

171 each with N = 10,000 individuals, where each individual comprised a vector of random values sampled

172 form a normal distribution, ~ $N(0,1)$, for a matrix, **X**, comprising 10 variables and matrix, **Y**, also

173 comprising 10 variables. In one population, **X** and **Y** were simulated with no linear association (the range

174 of "true" PLS correlations was 0.0459-0.0709 over 100 runs). In the other population, **X** and **Y** were

175 simulated with a fairly strong linear association (the range of true PLS correlations was 0.8239- 0.8340

176 over 100 runs). Within each simulation run, two sampling frames of n = 30 specimen numbers were

177 randomly generated, for 200 permutations. The first sampling frame was used to sample individuals from

178 each of the "no effect" population (no linear association) and "effect" population (linear association). In

179 each sample, r_{PLS} values were calculated and sampling distributions for r_{PLS} were generated, based on

180 1,000 random resampling permutations for the PLS analysis); $\hat{z}_{12}$ was calculated between samples, as in
181 Equation 2. For this comparison, we expected the null hypothesis to be rejected. Therefore, the proportion
182 of times it was rejected is a measure of statistical power for the difference in effects. (Likewise, the
183 proportion of times it was not rejected is a measure of the type II error rate.)

184 The second sampling frame was used to draw a second sample from each population, and r_{PLS}
185 values and their sampling distributions were calculated as above, but $\hat{z}_{12}$ values were calculated between
186 samples from the same population. For these comparisons, we expected the null hypothesis to be
187 supported, but in one case no integration (no effect) and in the other, substantial integration (effect)
188 should be found from samples of each population. Irrespective of the level of integration, this procedure
189 simulated multiple runs of a null model (no expected difference in integration between populations),
190 allowing us to ascertain whether the type I error rates – the proportion of times the null hypothesis was
191 rejected – was consistent in spite of different levels of morphological integration. In all cases, the
192 significance level for the null hypothesis was assigned as $\alpha = 0.05$. Type 1 error rates and statistical power
193 were calculated within runs as the proportion of 200 permutations that the null hypothesis was rejected.
194 Means and 95% confidence intervals were calculated across the 100 runs to evaluate the tendencies of
195 type I error rates and statistical power.

196 Additionally, because integration tests are often performed on geometric morphometric data, we
197 also considered the contribution of generalized Procrustes analysis (GPA) to generate inherently
198 correlated shape variables, and to what extent this would affect results of our proposed procedure.
199 Briefly, GPA scales anatomical landmark configurations to unit size, centers them, and rotates them via
200 least squares superimposition to render configurations invariant is size, position, or orientation. Procrustes
201 residuals– the aligned landmarks following GPA – are inherently correlated shape variables produced by
202 GPA. Therefore, even generating isotropic error on points of landmark configurations will produce
203 correlated variables within modules prior to measuring the PLS correlation between modules. We
204 considered the GPA-artifact by assigning the randomly generated data from the simulation experiment
205 above as residuals on mean configurations. We generated 5-point configurations of two-dimensional

landmarks (a vector of 10 values) in each simulation run. These configurations served as mean vectors, to which the row vectors of the previously generated random **X** and **Y** matrices were added as residuals to generate individual configurations within populations. For the "no effect" cases described above, we forced the 5-point configurations to be approximately uncorrelated ($r_{PLS} < 0.1$) between the mean configurations for **X** and **Y**. For the "effect" cases, mean configurations were the same. The mean configurations were also generated to have dispersion of several orders of magnitude greater than the random within-point dispersion, to ensure resulting individual landmark configurations were reasonable. In each permutation within each simulation run, GPA was performed and Procrustes residuals were used for r_{PLS} calculations. Thus, our results allowed us to evaluate both GM (Procrustes residuals) and non-GM (multivariate data) applications for consistently generated residuals, and whether GPA altered interpretations.

We repeated the entire process for 5 simulation runs with 20 sampling events and PLS analyses with 1,000 random permutations to consider whether the sampling distributions of r_{PLS} values were normally distributed. In each sampling event, we calculated a Shapiro-Wilk *W* statistic (i.e., 100 values total for each comparison), and the ranges of these values were qualitatively evaluated to determine if sampling distributions were approximately normally distributed (i.e., a Shapiro-Wilk *W* statistic that tends toward 1.0). All simulations were performed in R 3.2.0 (R Core Team 2015).

Results. The simulation experiment demonstrated that type I error rates were appropriate for both random multivariate data and Procrustes residuals (Fig. 2). For "no effect" comparisons, the mean type I error rates (0.0453 and 0.0494 for multivariate data and Prcrustes residuals, respectively) were approximately the same as the nominal significance level (0.05). Interestingly, the type I error rates were lower for comparisons between equal but large effects (i.e., when comparing datasets where both exhibited marked morphological integration of similar strength). A type I error was simulated in 0 and 1 permutation of the 100 × 200 total permutations, for multivariate data and Procrustes residuals, respectively. The nearly non-existent type I errors imply that when both datasets contain integration but their levels are similar, the method displays fewer false positives as compared to when comparing datasets

lacking integration (Fig.2). When comparing samples from the “no effect” and “effect” populations, the mean statistical power was 0.8133 and 0.883 for multivariate data and Procrustes residuals, respectively. Both of these values were, approximately equal to or greater than the generally desired value of 0.8 (Fig. 2), which represents a 4:1 trade-off between type II error risk and type I error risk (Cohen 1988). However, the inherently generated within-module correlations from GPA appear to slightly increase statistical power, perhaps owing to the additional correlation between mean configurations that was simulated. These results suggest that the two-sample test of integration disparity presented here behaves as expected, statistically, and GPA does not negatively impact results.

In terms of the appropriateness of the two-sample test of integration disparity for PLS analyses, we found no evidence to suggest that the sampling distributions of r_{PLS} statistics were non-normally distributed. For multivariate data, the ranges in *W* statistics for the “no effect” case (0.9822-0.9996) and “effect” case (0.9885-0.9997) were quite similar and sufficiently close to 1.0 in each case. For the Procrustes residuals, the ranges in *W* statistics for the “no effect” case (0.9910-0.9996) and “effect” case (0.9895-0.9996) were also quite similar and sufficiently close to 1.0 in each case. Thus, whether GPA was performed had no consequence, and any PLS analysis produced sufficiently normal sampling distributions.

It should be noted that the sampling distribution of r_{PLS} values is arbitrarily based on the *a priori* chosen number of PLS resampling permutations. One may wish to either choose a sufficiently large number of permutations or confirm that the standard deviation of the sampling distribution remains consistent under a range of permutations. For example, we found the standard deviation in our samples of 30 individuals in the simulation experiment was rather consistent between 200-10,000 random permutations, suggesting the 1,000 permutations we used was adequate for measuring the difference in strength of integration among datasets. We also found that when using a small number of PLS resampling permutations (e.g., 100-200), a large effect for the observed r_{PLS}, could skew the sampling distribution (which should be normally distributed). This problem is alleviated by simply removing the observed r_{PLS} from the sampling distribution, as a small-sample bias adjustment. For sufficiently large numbers of PLS

permutations (e.g., 1,000), this step was not needed, but also did not affect results. We, therefore, recommend removing the observed r_{PLS} as a procedural step to assure a more appropriate standard deviation of sampling distributions, especially for large effect sizes.

A Biological Example

To illustrate the method described above we conducted a comparison of levels of integration in cranial shape between rural and urban populations of the Mediterranean wall lizard *Podarcis muralis*. The data were part of a series of studies that evaluated the effects of urbanization on various aspects of phenotypic variation, including patterns of allometry, developmental stability, and integration in juvenile and adult lizards (see Lazic et al. 2015, 2016: available on Dryad and from the original authors). For this example, landmark-based geometric morphometric methods (Bookstein 1991; Mitteroecker and Gunz 2009; Adams et al. 2013) were used to characterize head shape, based on the positions of 28 homologous locations (Fig. 3A) collected from the dorsal view of 482 juvenile and 359 adult lizards from several localities. Of these, approximately half of the specimens were collected from rural locations (218 juveniles and 191 adults), and the remainder from urban sites (264 juveniles and 168 adults). For each specimen, a mirror image of their landmark locations was obtained by reflecting the coordinates about the mid-line, and a generalized Procrustes analysis was then performed to remove the effects of non-shape variation from the dataset (Fig. 3B). The symmetric component of shape was subsequently obtained by averaging landmark locations for each specimen and its mirror image. From the symmetric component of shape variation, we evaluated the degree to which the anterior and posterior regions of the head were integrated with one another. Landmarks were classified as anterior or posterior (Fig. 3A) based on the timing of ossification events during development (Lazic et al. 2015). For both juveniles and adults from rural and urban populations, the degree of integration between modules was estimated using partial least squares. Here the maximal covariation between modules was characterized using the PLS correlation (r_{PLS}), which was statistically evaluated using 1,000 random permutations. Additionally, *z*-scores were obtained for all four groups, and were statistically compared to one another using the procedure described

284 above. All analyses were performed in R 3.2.0 (R Core Team 2015) using the package *geomorph* (Adams
285 and Otárola-Castillo 2013; Adams et al. 2016), including the function, `compare.pls`, .which performs
286 the method introduced here.

287 *Results*. For both juvenile and adults from rural and urban lizard populations, the degree of
288 morphological integration between anterior and posterior regions of the head was large and highly
289 significant ($rural_{juv}$: r_{PLS} = 0.770, P_{rand} = 0.001; $rural_{adult}$: r_{PLS} = 0.826, P_{rand} = 0.001; $urban_{juv}$: r_{PLS} = 0.761,
290 P_{rand} = 0.001; $urban_{adult}$: r_{PLS} = 0.858, P_{rand} = 0.001). Generally, the observed integration of the frontal and
291 distal regions of the head was represented by a relative shortening of the snout accompanied by an
292 enlargement of the posterior area of the head (Fig. 3C); a pattern broadly observed in all groups. When
293 converted to effect sizes, all *z*-score values were very large (Fig. 3D), implying that the degree of
294 integration in each group greatly exceeded that which was expected by chance. Interestingly, when
295 compared using equation 2 above, we found no evidence of differences in levels of integration between
296 rural and urban juveniles or rural and urban adults, but significant differences existed in levels of
297 integration between juveniles and adults within each population; with adults displaying significantly
298 greater integration relative to juveniles (Table 1). Thus, while there was no evidence that environmental
299 disturbances have affected the strength of morphological integration in urban populations, the results
300 demonstrate that morphology is more integrated through ontogenetic time. Further, when combined with
301 the prior observation that morphological variation among adult specimens was reduced when compared to
302 that of juveniles (see Lazic et al. 2016), the pattern identified here is consistent with what is expected
303 under the hypothesis of developmental canalization (Hallgrímsson et al. 2002).
304

305 **Discussion**

306 An important question in evolutionary biology is whether different taxa or traits display similar
307 levels of morphological integration. Unfortunately, direct quantitative comparisons of the level of
308 integration across datasets have been hampered by the fact that the measures that characterize these
309 patterns (RV and r_{PLS}) are dependent on both sample size and the number of variables. Here we described

an unbiased effect size for quantifying the strength of morphological integration between sets of variables, utilizing partial least squares analysis and its permutation sampling distribution. We demonstrated that under the null hypothesis of a random association of variables, the approach displays a constant expected value, and the same variance, standard deviation, and confidence intervals across the entire spectrum of sample sizes and variable number. We further proposed a two-sample test to statistically evaluate the difference in effect sizes across two datasets. The approach displays appropriate type I error and statistical power, thereby providing a rigorous means of evaluating whether the degree of morphological integration differs between them. Thus the approach provides evolutionary morphologists with a consistent means of deciphering whether levels of integration are similar to one another in two or more datasets.

Extensions to the approach developed here can be envisioned that address a wider array of empirical challenges than the ones presented. For example, if the integration among three sets of variables is of biological interest, a three-block partial least squares approach may be utilized to characterize the degree of covariation among sets of variables (see Bookstein et al. 2003). Alternatively, one may evaluate the mean of the pairwise r_{PLS} values as a general test measure, as has been proposed for evaluating hypotheses of modularity (Klingenberg 2009; Adams 2016). In either case it is important to recognize that the method developed here simply provides a quantitative estimate on the magnitude, or strength of morphological integration among sets of variables. The method provides no description of the type of integration, or *how* traits covary with one another and in what manner. For this, understanding patterns of morphological integration and the set of coordinated shape changes it embodies must still be accomplished via a thorough examination of the singular vectors from the PLS and a visual inspection of the shape changes associated with the singular vectors (see Bookstein 2016 for discussion). Thus, a proper biological understanding of patterns of morphological integration is accomplished via the combination of a quantitative characterization and statistical assessment of the magnitude of integration, along with its anatomical interpretation.

Finally, while the biological concept of morphological integration has been embraced in

evolutionary biology for decades (Olson and Miller 1958), formal statistical tests of these patterns are still rather novel. As the theory of morphological integration develops, so too will the analytical methods for measuring integration, and their associated hypothesis tests (see e.g., Bookstein 2015). An important advance made here is that for any two measures of morphological integration – irrespective of the number of variables, number of specimens studied, or expected values in a null distribution – effect sizes calculated as standard deviates in sampling distributions are values that can be compared in a general two-sample hypothesis test. We chose to use the correlation coefficient from two-block PLS as the basis for this test, but one could have likewise used maximum singular values, or vector correlations (or angles) between left and right singular vectors found through PLS, or other statistically-relevant summaries. Since standard deviates can be calculated using any test statistic with a sampling distribution – essentially any statistic if resampling experiments are used – the hypothesis testing framework proposed here is merely a methodological extension of two sample *Z*-tests, but using resampling experiments to generate sampling distributions rather than requiring *a priori* knowledge of population standard deviations. Thus, as new analytical approaches are developed for evaluating patterns of morphological integration, the test procedure described here may be utilized for comparing the degree of integration observed in different datasets based on those measures. Ultimately however, the choice of which test statistic to utilize in this procedure must be driven by biology. In the case of morphological integration, biological interpretations of such patterns depend on a deep understanding of how covariation patterns are embodied in terms of their singular vectors (see Bookstein 2016). Nevertheless, as the field of evolutionary morphological integration evolves, and better conceptual measures of morphological integration are developed, a hypothesis-testing framework developed here is already in place, and ready for such advances.

Acknowledgments

We thank A. Kaliontzopoulou for her comments on the manuscript. A Kaliontzopoulou and M. Lazić kindly provided the data for the empirical example and the image for Fig. 2a. This work was sponsored in part by National Science Foundation grants DEB-1556379 (to DCA) and DEB-1556540 (to

362 MLC).

References

Adams, D. C. 2016. Evaluating modularity in morphometric data: challenges with the RV coefficient and a new test measure. Methods Ecol. Evol. 7:565-572.

Adams, D. C., M. Collyer, and E. Sherratt. 2016. geomorph 3.0.1: Software for geometric morphometric analyses. R package version 3.0.1. http://CRAN.R-project.org/package=geomorph.

Adams, D. C. and R. N. Felice. 2014. Assessing trait covariation and morphological integration on phylogenies using evolutionary covariance matrices. PLoS ONE 9:e94335.

Adams, D. C. and E. Otárola-Castillo. 2013. geomorph: an R package for the collection and analysis of geometric morphometric shape data. Methods Ecol. Evol. 4:393-399.

Adams, D. C., F. J. Rohlf, and D. E. Slice. 2013. A field comes of age: Geometric morphometrics in the 21st century. Hystrix 24:7-14.

Armbruster, W. S., C. Pelabon, G. H. Bolstad, and T. F. Hansen. 2014. Integrated phenotypes: understanding trait covariation in plants and animals. Phil. Trans. Roy. Soc. London B. 369:20130245.

Bookstein, F. L. 1991. Morphometric tools for landmark data: geometry and biology. Cambridge University Press, Cambridge.

Bookstein, F. L. 2015. Integration, disintegration, and self-similarity: characterizing the scales of shape variation in landmark data. Evol. Biol. 42:395-426.

Bookstein, F. L. 2016. The inappropriate symmetries of multivariate statistical analysis in geometric morphmetrics. Evol. Biol. 43:DOI 10.1007/s11692-11016-19382-11697.

Bookstein, F. L., P. Gunz, P. Mitteroecker, H. Prossinger, K. Schaefer, and H. Seidler. 2003. Cranial integration in Homo: singular warps analysis of the midsagittal plane in ontogeny and evolution. J. Hum. Evol. 44:167-187.

Cheverud, J. M. 1982. Phenotypic, genetic, and environmental morphological integration in the cranium. Evolution 36:499-516.

Cheverud, J. M. 1996. Developmental integration and the evolution of pleiotropy. Am. Zool. 36:44-50.

389 Cohen, J. 1988. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum, Hillsdale, New
390 Jersey. .
391 Drake, A. G. and C. P. Klingenberg. 2010. Large scale diversification of skull shape in domestic dogs:
392 disparity and modularity. Am. Nat. 175:289-301.
393 Escoufier, Y. 1973. Le traitement des variables vectorielles. Biometrics 29:751–760.
394 Fruciano, C., P. Franchini, and A. Meyer. 2013. Resampling-Based Approaches to Study Variation in
395 Morphological Modularity. PLoS ONE 8:e69376.
396 Gómez-Robles, A. and P. D. Polly. 2012. Morphological integration in the hominin dentition:
397 evolutionary, developmental, and functional factors. Evolution 66:1024-1043.
398 Goswami, A. and P. D. Polly. 2010. Methods for studying morphological integration and modularity. Pp.
399 213-243 *in* J. Alroy, and G. Hunt, eds. Quantitative Methods in Paleobiology.
400 Goswami, A., J. B. Smaers, C. Soligo, and P. D. Polly. 2014. The macroevolutionary consequences of
401 phenotypic integration: from development to deep time. Phil. Trans. Roy. Soc. London B.
402 369:20130254.
403 Hallgrímsson, B., K. Willmore, and B. K. Hall. 2002. Canalization, developmental stability, and
404 morphological integration in primate limbs. Am. J. Phys. Anthropol. 119:131–158.
405 Klingenberg, C. P. 2008. Morphological integration and developmental modularity. Ann. Rev. Ecol. Evol.
406 Syst. 39:115-132.
407 Klingenberg, C. P. 2009. Morphometric integration and modularity in configurations of landmarks: tools
408 for evaluating a priori hypotheses. Evol. Develop. 11:405-421.
409 Klingenberg, C. P. 2014. Studying morphological integration and modularity at multiple levels: concepts
410 and analysis. Phil. Trans. Roy. Soc. London B. 369:20130249.
411 Lazic, M. M., M. A. Carretero, J. Crnobrnja-Isailovic, and A. Kaliontzopoulou. 2015. Effects of
412 environmental disturbance on phenotypic variation: an integrated assessment of canalization,
413 developmental stability, modularity, and allometry in lizard head shape. Am. Nat. 185:44-58.

414 Lazic, M. M., M. A. Carretero, J. Crnobrnja-Isailovic, and A. Kaliontzopoulou. 2016. Postnatal dynamics
415 of develpomental stability and canalization of lizard head shape under different environmental
416 conditions. Evol. Biol. DOI 10.1007/s11692-016-9377-4.
417 Magwene, P. M. 2001. New tools for studying integration and modularity. Evolution 55:1734–1745.
418 Márquez, E. J. 2008. A statistical framework for testing modularity in multidimensional data. Evolution
419 62:2688–2708.
420 Martin-Serra, A., B. Figueirido, J. A. Perez-Claros, and P. Palmqvist. 2015. Patterns of morphological
421 integration in the appendicular skeleton of mammalian carnivores. Evolution 69:321-340.
422 Mitteroecker, P. and F. L. Bookstein. 2007. The conceptual and statistical relationship between
423 modularity and morphological integration. Syst. Biol. 56:818–836.
424 Mitteroecker, P. and P. Gunz. 2009. Advances in geometric morphometrics. Evol. Biol. 36:235-247.
425 Mitteroecker, P., P. Gunz, M. Bernhard, K. Schaefer, and F. L. Bookstein. 2004. Comparison of cranial
426 ontogenetic trajectories among great apes and humans. J. Hum. Evol. 46:679-698.
427 Monteiro, L. R., V. Bonato, and S. F. d. Reis. 2005. Evolutionary integration and morphological
428 diversification in complex morphological structures: Mandible shape divergence in spiny rats
429 (Rodentia, Echimyidae). Evol. Develop. 7:429-439.
430 Neaux, D., E. Gilissen, W. Coudyzer, and F. Guy. 2015. Integration Between the Face and the Mandible
431 of Pongo and the Evolution of the Craniofacial Morphology of Orangutans. Am. J. Phys. Anthropol.
432 158:475-486.
433 Olson, E. C. and R. L. Miller. 1958. Morphological Integration. University of Chicago Press, Chicago.
434 Pigliucci, M. 2003. Phenotypic integration: studying the ecology and evolution of complex phenotypes.
435 Ecol. Lett. 6:265-272.
436 R Core Team. 2015. R: a language and environment for statistical computing. Version 3.2.0.
437 http://cran.R-project.org. R Foundation for Statistical Computing, Vienna.
438 Rohlf, F. J. and M. Corti. 2000. The use of partial least-squares to study covariation in shape. Syst. Biol.
439 49:740-753.

440 Smilde, A. K., H. A. L. Kiers, S. Biklsma, C. M. Rubingh, and M. J. vanErk. 2009. Matrix correlations
441 for high-dimensional data: the modified RV-coefficient. Bioinformatics 25:401-405.
442 Young, R. L. and A. V. Badyaev. 2006. Evolutionary persistence of phenotypic integration: influence of
443 developmental and functional relationships on complex trait evolution. Evolution 60:1291–1299.
444

Table 1. Results from empirical example comparing levels of morphological integration in juvenile and adult lizards from urban and rural populations. A) Matrix of pairwise differences in PLS effect sizes, and B) their associated significance levels. Biologically-relevant focal comparisons are underlined; significant focal comparisons are shown in bold. Populations are designated as: UR: urban, RU: rural, Ad: adult, Juv: juvenile.

Z	UR_{Ad}	RU_{Ad}	UR_{Juv}	RU_{Juv}		P	UR_{Ad}	RU_{Ad}	UR_{Juv}	RU_{Juv}
UR_{Ad}	0					UR_{Ad}	0			
RU_{Ad}	0.105	0				RU_{Ad}	0.459	0		
UR_{Juv}	2.818	2.777	0			UR_{Juv}	**0.002**	0.002	0	
RU_{Juv}	2.447	2.399	0.296	0		RU_{Juv}	0.007	**0.008**	0.383	0

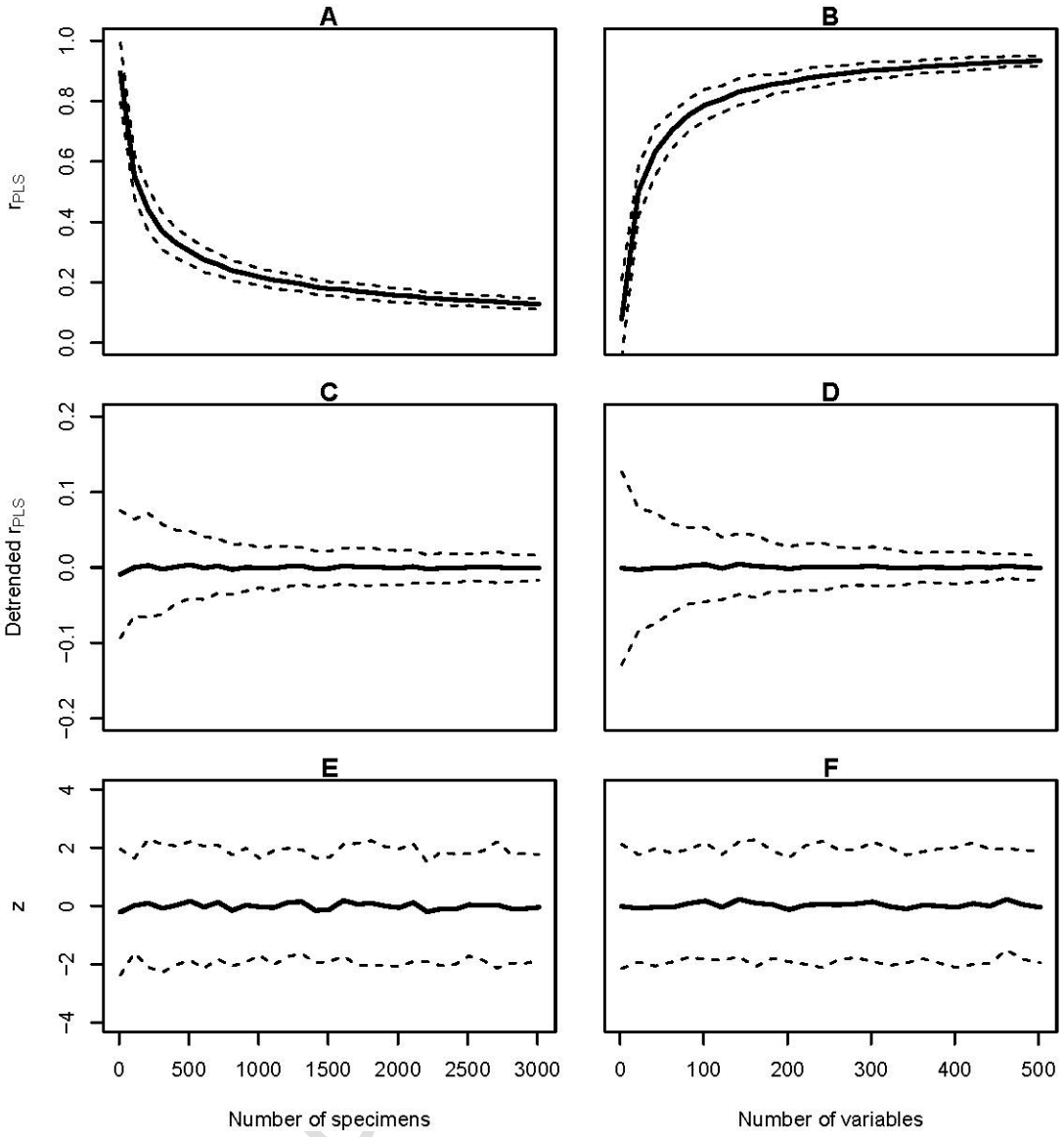

Fig. 1. Evaluation of r_{PLS} under the hypothesis of random associations of variables. Mean and 95% confidence intervals of RV values obtained from (A) 100 datasets simulated across a range of sample sizes, and from (B) 100 datasets simulated across a range of variable number. Mean and 95% confidence intervals of detrended r_{PLS} for the same simulations of (C) 100 datasets simulated across a range of sample sizes, and from (D) 100 datasets simulated across a range of variable number. Mean and 95% confidence intervals of *z*-scores for the same simulations of (E) 100 datasets simulated across a range of sample sizes, and from (F) 100 datasets simulated across a range of variable number.

.

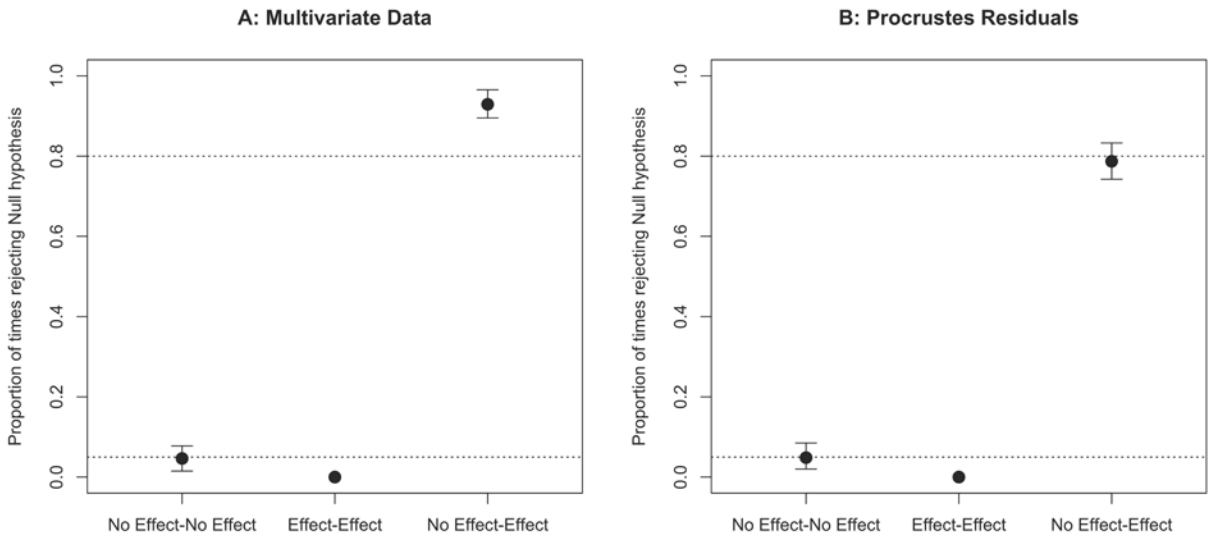

Fig. 2. Proportion of times the null hypothesis was rejected in 100 runs of 200 comparisons of morphological integration for (A) multivariate data and (B) Procrustes residuals. Samples of size, n = 30 were obtained from populations (N = 10,000) with no integration (no effect) and substantial integration (effect). Means with 95% confidence limits are shown, unless the proportion was rather invariant, in which case, error bars are not included. Dotted lines are shown for an expected type I error rate of 0.05 or a statistical power of 0.80; within-population comparisons simulate type I errors (first two) and between-population comparisons simulate statistical power (last).

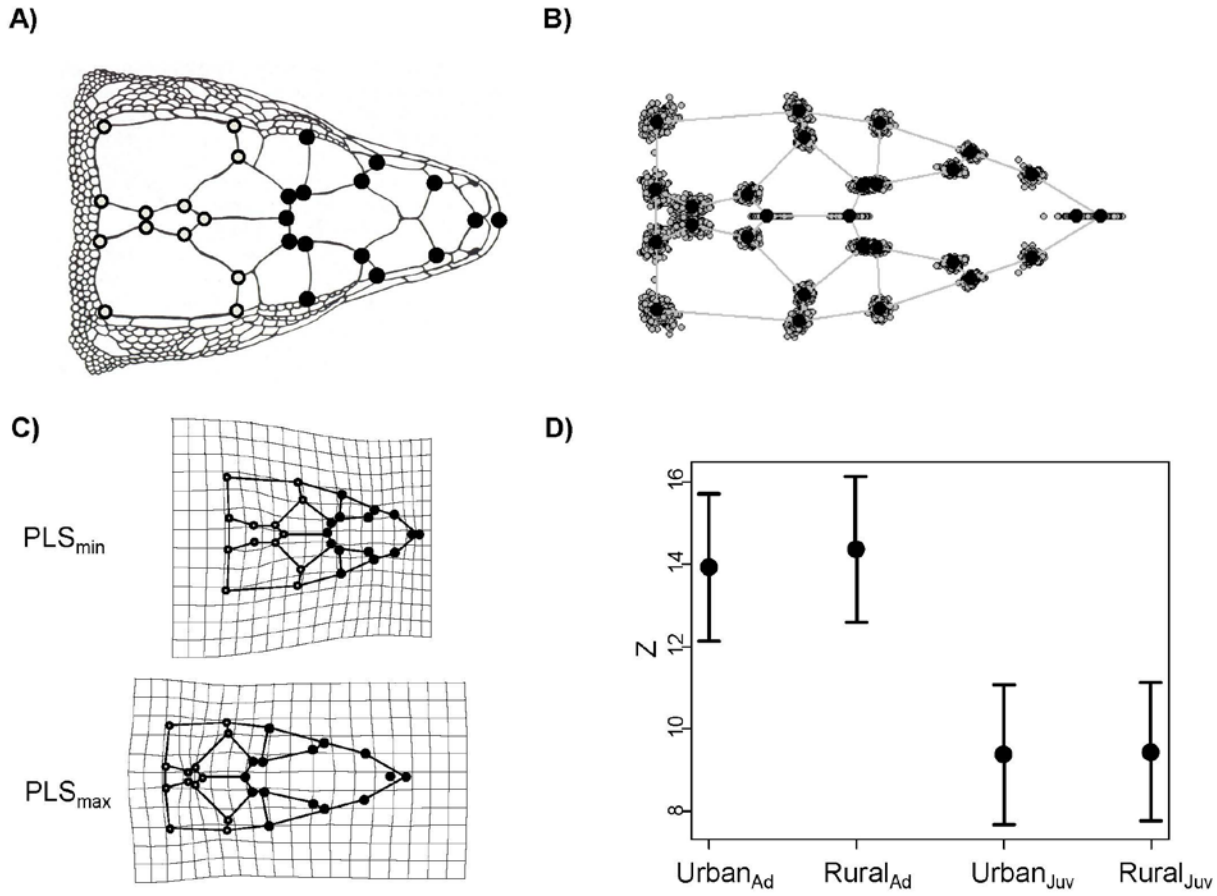

Fig. 3. Graphical summary of results from the empirical example. (A) Locations of 28 landmarks on the dorsal view of a lizard head. Landmarks from the anterior module are designated by open circles while those from the posterior module are designated by closed circles (from Lazic et al. 2015). (B) Procrustes superimposition of all specimens (gray) with the mean specimen shown in black. (C) Thin-plate spline deformation grids representing specimens at the extremes of the PLS axis for the adult rural population. Deformation grids are accentuated by a factor of two to facilitate visual interpretation. (D) Levels of integration in juveniles and adults from both rural and urban populations shown as *z*-scores and their 95% confidence intervals obtained from the standard error of the permutation sampling distributions for each group.